Argumentative Forecasting

Extended Abstract

Benjamin Irwin
Imperial College London
London, UK
benjamin.irwin19@imperial.ac.uk

Antonio Rago
Imperial College London
London, UK
a.rago@imperial.ac.uk

Francesca Toni
Imperial College London
London, UK
ft@imperial.ac.uk

ABSTRACT

We introduce the *Forecasting Argumentation Framework (FAF)*, a novel argumentation framework for forecasting informed by recent judgmental forecasting research. FAFs comprise update frameworks which empower (human or artificial) agents to argue over time with and about probability of scenarios, whilst flagging perceived irrationality in their behaviour with a view to improving their forecasting accuracy. FAFs include three argument types with future forecasts and aggregate the strength of these arguments to inform estimates of the likelihood of scenarios. We describe an implementation of FAFs for supporting forecasting agents.

KEYWORDS

Argumentation; Forecasting; Multi-Agent Debate

ACM Reference Format:
Benjamin Irwin, Antonio Rago, and Francesca Toni. 2022. Argumentative Forecasting: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

Historically, humans have performed inconsistently in judgemental forecasting [15, 20], which incorporates subjective opinion and probability estimates to predictions [14]. Yet, human judgement remains essential when pure statistical methods are inapplicable, e.g. where historic data alone is insufficient or for one-off, 'unknowable' events [1, 3, 18]. Effective tools to help humans improve their predictive abilities thus have enormous potential for impact.

Research on judgemental forecasting (see [14, 23] for overviews) is instructive in establishing the desired properties for systems for supporting forecasting. In addition to reaffirming the importance of fine-grained probabilistic reasoning [16], this literature points to the benefits of some group techniques versus solo forecasting [13, 21], of synthesising qualitative and quantitative information [14], of combating agents' irrationality [8] and of high agent engagement with the forecasting challenge, e.g. robust debating [13] and frequent prediction updates [16].

Computational argumentation (see [2, 4] for overviews), involves reasoning with uncertainty and resolving conflicting information and as such is an ideal candidate for aggregating the broad, polymorphous set of information involved in judgemental group forecasting. Subsets of the requirements for forecasting systems are addressed by individual formalisms, e.g. *probabilistic argumentation* [9–12, 22] may effectively represent and analyse uncertain arguments about the future. However, we posit that a purpose-built argumentation framework for forecasting is essential to effectively utilise computational argumentation's reasoning capabilities in this context. We draw from forecasting literature to inform the design of a new computational argumentation framework tailored to forecasting.

2 A BIRD'S EYE VIEW

Here we describe informally our argumentative forecasting process. This starts with an initial *forecast* $\mathcal{F}$ on the probability of an *outcome* and progressively updates this forecast as a result of independent debates between agents involved in the forecasting. Each debate is based on a *proposal argument* $\mathcal{P}$ for revising $\mathcal{F}$, taking into account some new and relevant evidence or context as captured by *amendment arguments* proposing to *increase* or *decrease* the forecast in the proposal argument, as well as *pro/con arguments* (borrowed from QuAD frameworks [5]). The latter can only be in relation with amendment arguments (thus leading to hierarchically structured debates with proposal arguments at the top, amendment arguments in the middle, and pro/con arguments at the bottom, possibly in several strata). After each debate, the current forecast is updated to form the next forecast. The nature of this update — whether it should *increase* or *decrease* and by how much – is determined argumentatively, following the agents' *voting* on the pro/con arguments, their *individual forecasts* and the verification that they are *rational* given their *confidence* in the proposal argument. Thus we understand an agent's (ir)rationality, informally, as the degree to which its probabilistic forecast (dis)agrees with the votes it has put forward on the arguments. The aggregation of all (rational) individual forecasts becomes the current forecast of the group. This may be updated subsequently via the same process, if time allows. The process can be summarised as follows:

Step 1: An agent opens the debate by introducing $\mathcal{P}$, which proposes a revision of $\mathcal{F}$.

Step 2: A shared debate develops as agents add increase/decrease amendment arguments ($\mathcal{X}^{\uparrow}/\mathcal{X}^{\downarrow}$) as well as pro/con arguments ($\mathcal{X}^{+}/\mathcal{X}^{-}$) debating the reasoning behind $\mathcal{X}^{\uparrow}$ and $\mathcal{X}^{\downarrow}$. $\mathcal{X}^{+}$ and $\mathcal{X}^{-}$ arguments are voted on by the agents.

Step 3: All agents advance an individual forecast.

Step 4: Irrational forecasting is prevented by stopping individual agents from making forecasts which run contrary to their *confidence score*, which encodes their belief *wrt* $\mathcal{P}$. An agent's *confidence score* is calculated from their votes using a variant of the DF-QuAD algorithm [19] originally defined for determining arguments' dialectical strength in QuAD frameworks.

Step 5: An aggregated forecast is produced from all the agents' (rational) forecasts, using a weighted mean function which adjusts

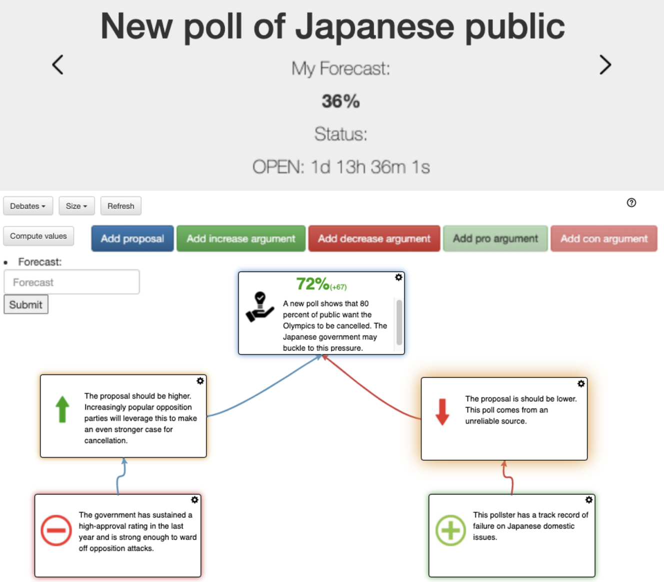
Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

each user's influence based on their previous forecasting accuracy (Brier Score [7]). This new aggregated forecast becomes $\mathcal{F}$.

Steps 1-5 are repeated up to a preset time limit, at which point the aggregated forecast is returned.

Update frameworks underpin Steps 1-3 and are a novel form of computational argumentation framework. They amount to proposal, amendment and pro/con arguments and relations between them (in the spirit of QuAD frameworks [5]). We stress that the connection with computational argumentation goes past the use of (various types of) arguments and relations between them. Indeed, the ultimate acceptability of a proposal argument is encoded in its confidence score, which is calculated by aggregating the DF-QuAD scores [19] of the amendment arguments (in the individual agent's view of the update framework).

A *Forecasting Argumentation Framework (FAF)* is made up of sequences of these *update frameworks*, addressing a single forecasting question over time. Note that FAFs can be seen as forming the basis for *deliberative democratic* [6] forecasting. $\mathcal{F}$ is initialised with a pre-agreed 'base-rate' forecast, e.g. based on historic data. Subsequently, the forecast is revised by one or more (non-concurrent) debates, opened and resolved by participating agents within the time limit. The composite nature of this process enables the appraisal of new epistemic and probabilistic contexts as and when they arise. Rather than confronting an unbounded forecasting question with a diffuse set of possible debates open at once, all agents concentrate their argumentation on a single topic (a proposal) at any given time.


New poll of Japanese public

My Forecast:

36%

Status:

OPEN: 1d 13h 36m 1s

- Forecast:

72%(+67)

A new poll shows that 80 percent of public want the Olympics to be cancelled. The Japanese government may buckle to this pressure.

The proposal should be higher. Increasingly popular opposition parties will leverage this to make an even stronger case for cancellation.

The proposal is should be lower. This poll comes from an unreliable source.

The government has sustained a high-approval rating in the last year and is strong enough to ward off opposition attacks.

This pollster has a track record of failure on Japanese domestic issues.

Figure 1: An example update framework in Arg&Forecast, where $\mathcal{P}$ is represented as the root node, $\mathcal{X}^{\uparrow}$ and $\mathcal{X}^{\downarrow}$ are marked with up/down arrows respectively and $\mathcal{X}^{+}$ and $\mathcal{X}^{-}$ arguments are marked with plus/minus icons, respectively.

3 IMPLEMENTATION: ARG&FORECAST

We produced a preliminary implementation of FAFs in the form of a publicly available web platform[1] called *Arg&Forecast*, extending the open source platform *Arg&Dec* [17][2]. Arg&Forecast enables users to initialise FAFs and debate, in real time, in a series of update frameworks, aided by a visual graph interface (illustrated in Figure 1, where the overall forecasting question is 'Will the Tokyo Olympics be cancelled/postponed to another year?').

Users can initialise a FAF and then an update framework within that FAF by introducing a proposal node to revise $\mathcal{F}$ (Step 1 in Section 2). In the example in Figure 1, $\mathcal{P}$ contains new evidence relevant to the forecasting question, namely the emergence of polling data showing that the Japanese public would like the Olympics to be cancelled. Accordingly, $\mathcal{P}$ suggests a new forecast reflecting increased probability of cancellation (72%). Subsequently, and in real time, other participants argue about $\mathcal{P}$ (Step 2), by adding the amendment and pro/con arguments in Figure 1, using the panel atop the graph interface. Concurrently, participating users can vote on the pro/con arguments by selecting the options icon at the top right corner of the nodes holding the pro/con arguments. Each user may give each argument a personal base score (in [0,1]) and, in turn, this impacts their overall confidence score, calculated by *Arg&Forecast.* When users advance an individual forecast using the forecast input box on the left (Step 3), this confidence score is used to flag irrationality (Step 4) by blocking offending forecasts and alerting users with an onscreen dialogue. The update framework can close when a) its time limit set at the outset runs out or b) the debate has stabilised, and all users have made rational forecasts (Step 5). The final aggregated forecast of the latest update framework becomes the current forecast for the overall question. At this point, debate participants are free to add further update frameworks based on new evidence or information. This cycle continues until the question runs beyond its initial preset time window, or the debate is resolved by the revelation of a ground truth.

Evaluation. We explored, with a small scale human user study of *Arg&Forecast*, possible strengths of the platform in facilitating higher forecasting accuracy and debate engagement. In a real time forecasting challenge, a subgroup which used *Arg&Forecast* significantly outperformed (21% lower overall Brier score) a control subgroup, who used a chatroom variation on the standard focus group forecasting methodology [13]. The platform group also carried out 151% more debate engagements (proposal, amendment, pro/con arguments) and 57% more forecasts.

4 FUTURE WORK

There are a multitude of possible directions for future work, including: (i) considering multi-valued outcomes (e.g. 'Who will win the next UK election?'), (ii) integrating further rationality constraints, (iii) constraining agents' argumentation (e.g. by using past Brier scores to limit the quantity or strength of agents' arguments), and (iv) conducting further experiments with Arg&Forecast or with other implementations of our framework.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Anthony Hunter for his helpful contributions to discussions in the build up to this work.

[1] https://continuousargumentframework.com/

[2] https://github.com/dariopellegrini/arganddec

REFERENCES

[1] Meysam Arvan, Behnam Fahimnia, Mohsen Reisi, and Enno Siemsen. 2019. Integrating human judgement into quantitative forecasting methods: A review. *Omega (United Kingdom)* 86 (2019). https://doi.org/10.1016/j.omega.2018.07.012

[2] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. 2017. Towards Artificial Argumentation. *AI Magazine* 38, 3 (2017), 25–36.

[3] Shari De Baets and Nigel Harvey. 2020. Using judgment to select and adjust forecasts from statistical models. *European Journal of Operational Research* 284 (2020). Issue 3. https://doi.org/10.1016/j.ejor.2020.01.028

[4] Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre (Eds.). 2018. *Handbook of Formal Argumentation.* College Publications.

[5] Pietro Baroni, Marco Romano, Francesca Toni, Marco Aurisicchio, and Giorgio Bertanza. 2015. Automatic evaluation of design alternatives with quantitative argumentation. *Argument and Computation* 6 (2015). Issue 1. https://doi.org/10.1080/19462166.2014.1001791

[6] Joseph M. Bessette. 1980. Deliberative democracy: The majority principle in republican government. In *How Democractic is the Constitution?*, R.A. Goldwin and W.A. Schambra (Eds.). American Enterprise Institute for Public Policy Research, Washington, 102–116.

[7] Glenn W. Brier. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* 78 (1950). Issue 1. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

[8] Welton Chang, Eva Chen, Barbara Mellers, and Philip E. Tetlock. 2016. Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making* 11 (2016). Issue 5.

[9] Bettina Fazzinga, Sergio Flesca, and Filippo Furfaro. 2018. Probabilistic bipolar abstract argumentation frameworks: Complexity results. *IJCAI International Joint Conference on Artificial Intelligence* 2018-July. https://doi.org/10.24963/ijcai.2018/249

[10] Anthony Hunter. 2013. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning* 54 (2013). Issue 1. https://doi.org/10.1016/j.ijar.2012.08.003

[11] Anthony Hunter, Sylwia Polberg, and Matthias Thimm. 2020. Epistemic graphs for representing and reasoning with positive and negative influences of arguments. *Artificial Intelligence* 281 (2020). https://doi.org/10.1016/j.artint.2020.103236

[12] Anthony Hunter and Matthias Thimm. 2014. Probabilistic argumentation with incomplete information. *Frontiers in Artificial Intelligence and Applications* 263. https://doi.org/10.3233/978-1-61499-419-0-1033

[13] Jon Landeta, Jon Barrutia, and Aitziber Lertxundi. 2011. Hybrid Delphi: A methodology to facilitate contribution from experts in professional contexts. *Technological Forecasting and Social Change* 78 (2011). Issue 9. https://doi.org/10.1016/j.techfore.2011.03.009

[14] Michael Lawrence, Paul Goodwin, Marcus O'Connor, and Dilek Önkal. 2006. Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting* 22 (2006). Issue 3. https://doi.org/10.1016/j.ijforecast.2006.03.007

[15] Spyros Makridakis, Robin M. Hogarth, and Anil Gaba. 2010. Why forecasts fail. What to do instead. *MIT Sloan Management Review* 51 (2010). Issue 2.

[16] Barbara Mellers, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, and Philip E. Tetlock. 2015. Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science* 10 (2015). Issue 3. https://doi.org/10.1177/1745691615577794

[17] Dario Pellegrini. 2015. arganddec. https://github.com/dariopellegrini/arganddec. [Online; accessed 15-September-2021].

[18] Fotios Petropoulos, Robert Fildes, and Paul Goodwin. 2016. Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour? *European Journal of Operational Research* 249 (2016). Issue 3. https://doi.org/10.1016/j.ejor.2015.06.002

[19] Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. 2016. Discontinuity-Free Decision Support with Quantitative Argumentation Debates. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016.* 63–73. http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12874

[20] Philip E. Tetlock. 2017. *Expert political judgment: How good is it? How can we know?* Princeton University Press.

[21] Philip E. Tetlock, Barbara A. Mellers, Nick Rohrbaugh, and Eva Chen. 2014. Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate. *Current Directions in Psychological Science* 23 (2014). Issue 4. https://doi.org/10.1177/0963721414534257

[22] Matthias Thimm. 2012. A probabilistic semantics for abstract argumentation. *Frontiers in Artificial Intelligence and Applications* 242. https://doi.org/10.3233/978-1-61499-098-7-750

[23] Maximilian Zellner, Ali E. Abbas, David V. Budescu, and Aram Galstyan. 2021. A survey of human judgement and quantitative forecasting methods. Issue 2. https://doi.org/10.1098/rsos.201187